7th International Conference on Seismology & Earthquake Engineering
18-21 May 2015

MOVING TOWARD LESS UNCERTAINTY SEISMIC RISK PREDICTION USING GRANULAR COMPUTING ALGORITHM

H. S. ALINIA

M.Sc. Graduate, GIS Division, Department of Surveying and Geomatics Engineering, College of Eng., University of Tehran, Tehran, Iran,h_s_Alinia@yahoo.com

M. R. DELAVAR

Center of Excellence in Geomatics Eng. and Disaster Management, Dept. of Surveying and Geomatics Eng., College of Eng., University of Tehran, Tehran, Iran, mdelavar@ut.ac.ir

M. ZARE

Seismology Research Center, International Institute of Earthquake Engineering and Seismology (IIEES), Tehran, Iran, mzare@iiees.ac.ir

A. MOHSENI

Research Associate, Seismology Research Center, International Institute of EarthquakeEngineering and Seismology (IIEES), Tehran, Iran,Email: a.mohseni@iiees.ac.ir

Keywords: Data Mining, Uncertainty, Earthquake

ABSTRACT

Iran is one of the seismically active areas of the world due to its position in the Alpine-Himalayan mountain system. So, strong earthquakes in this area have caused a high toll of casualties and extensive damage over the last centuries.

Pre-determining locations and intensityof seismic area of a city is considered as a complicateddisaster management problem. As, this problem generally depends on various criteria, one of the most important challenges concerned is the existence of uncertainty regarding inconsistency in combining influencing criteria and extracting more consistent knowledge forthe next predictions. To overcome this problem, this paper proposes a new approach for seismic risk knowledge discovery based on granular computing theory. One of the significant properties of this method is inductionof more compatible rules having zero inconsistency fromexisting databases. Furthermore, in this approach non redundant covering rules will be extracted for consistent classification where one object maybe classified with two or more non-redundant rules.

In this paper, the seismic risk of the area between 58˚ 24' E, 60˚ 24' E Longitude and 27˚ 45' N, 29˚ 25' N Latitude around occurred near Reygan (Kerman Province), South-East of Iran where a devastating earthquake happened is considered as the study area. The result of this paper exhibits why granular computing is proposed to decrease the uncertainty of knowledge extracted from input large dataset.

INTRODUCTION

On December 26, 2003 when a major earthquake with magnitude 6.6 Mw hit the south-eastern of Iran, Kerman province, at 05:26AM (Iran standard time), the area most affected was the city of Bam where more than 43,000 people were killed, an estimated 30,000 injured and up to 75,000 left homeless, according to official estimates. Seven years after, another deadly earthquake struck its neararea. The epicentre of this event that was happened at 22:11(local time) on December 20, 2010 with magnitude 6.2 has been located at 28.36N, 59.22E (IIEES)approximately 52 kilometer of south of Khavari inMohammad-abadReygan. According to reports, 1150 residential area around Fahraj village were damaged nearly 30 to 70 percent. To cope with such natural disaster, is a big challenge for scientist and researches around the globe. Although the resulting devastation of this natural disaster can be mitigated through effective disaster management strategies such as by pre-disaster risk assessment and effective post-disaster response, it is impractical to eliminate it or predict it in a long time to prepare people to confront. It is obvious that the earthquakeis considered as a very complicated natural disaster and depends on various criteria. So, one of the significant challenges concerned is the existence of uncertainty to classify objects based on their similarities.This paper concentrates on using granular computing theory to study earthquake and its effective factors in various aspects specifically constructing granules in which each memberhas similar attribute-value by implementing a proposed computational method over granules. This paper proposes an algorithm for mining task to search for most suitable with less uncertain knowledge from earthquake information table constructed by some effective criteria and their values.

As an emerging field of study, Granular Computing (GrC) is both new and old. On the one hand, the term “granular computing” was first suggested by T.Y. Lin in 1997. Indeed, Granular computing is an umbrella term to cover any theories, methodologies, techniques, and tools that make use of granules in problem solving.

Concept of granular computing has applied by many authors to re-examine many classic problems, in order to obtain new understandings and more insights. It is a study of a general theory of problem solving based on different levels of granularity and detail. Also, the basic ideas of granular computing, i.e., problem solving with different granularities, have been explored in many fields, such as, interval analysis, rough set theory, in problem solving, Dempster-Shafer theory of evidence and many others. Also this concept was used in assessment of seismic vulnerability of Tehran, Capital of Iran.

Data mining is one of such problems explored by some researchers. This paper is another attempt to exhibit how granular computing approach is used for data mining. It serves dual purposes: demonstrating the potential of granular computing on one hand and exploring a new perspective of data mining in the seismic risk framework.

OVERVIEW OF GRANULAR COMPUTING

The basic ideas and principles of granular computing are not entirely new and have indeed been investigated in many disciplines of social and natural sciences. The study of granular computing aims at arriving at a new powerful philosophical view and a general problem-solving theory. They are referred to as structured thinking and problem-solving.

Two essential tasks in data mining are the representation of objects and the identification of forms and types of knowledge to be mined. Broadly, granular computing can be studied based on the notions of representation and process, which were also used by Marr in the study of vision. The representation concerns granules and their organizations in terms of levels, networks, and hierarchies. One focuses on common features and universally applicable principles for the understanding, description, organization, and formulation of various problems across many different disciplines. The process deals with (computational) methods that manipulate granules and granular structures. Many computational operations can be performed on granules such as reasoning, inferencing and learning. Elements in a granule are grouped together by indistinguishability, similarity, proximity or functionality. Algorithms for a granulation are a kind of semantic interpretation of why two objects are put into the same granule and how two objects are related with each other. Also, relationships between granules which can be interpreted as ordering, closeness, dependency or association between granules determine the linkage of the granules.

This paper exemplified data mining especially rule-based mining in two steps; the formation of concepts and the identification of relationship between concepts. Formal concept analysis may be considered as a concrete model of granular computing. It deals with the characterization of a concept by a unit of thoughts consistsof the intension and the extension of the concept.

From the stand point of granular computing in this research, the concept of seismic risk assessment can be exemplified at two parts, extension, i.e., a set of objects as instances of pre-species category of seismic risk and intension which consists of all properties or attributes with more effective impacts in occurrence of earthquake, that are valid for all those areas where the concept applies. In this approach, each object is represented by the values of a set of attributes and the knowledge mined from a sample dataset is illustrated in the form of rules.

INFORMATION TABLE

Information tables are used as a basic knowledge in granular computing models. It represents all available information and knowledge in a form of relationship between objects and value of their attributes. That is, objects are only perceived, observed, or measured by using a finite number of properties.

Definition of an information table is shown in the following tuple:

$$S = (U, At, L, \{V_a | a \in At\}, \{I_a | a \in At\}) \quad (1)$$

where, U is a finite non-empty set of objects,
At is a finite non-empty set of attributes,
L is a language defined by using attributes in At,
Va is a non-empty set of values ofa ∈At,
Ia: U → Vais an information function that maps an object of U to exactly one possible value of attribute "a" in "Va".

By considering an attribute a ∈At, an object x ∈U takes only one value from the domain Va of a. Let a(x) = Ia(x) denote the value of x on a. So, for an attribute a ∈ At and x, y ∈ U, an equivalence relation Ea is given by:

$xE_a y \Leftrightarrow a(x) = a(y)$

With respect to all attributes in A, x and y are indiscernible, if and only if they have the same value for every attribute in A. Therefore, a language L is defined for describing objects of the universe in an information table. The decision logic language (DL-language) studied by Pawlak is adopted. This logic language is defined for the information table to provide formal descriptions of various notions.

The information table of this research is constructed by existing knowledge from 18 seismic areas around Mohammad-abadReygan(Figure 1) in the historical events, arranged in 18 rows of objects and 6 columns of attributes. The six attributes which are considered in this paper include: the rate of seismic risk of the area, the existence of fault in the area and its length, fault coincidence in that area, occurrence of historical earthquake with magnitude more than 6 Richter and occurrence of historical earthquake with magnitude between 4 and 6 Richter. It is assumed that there is a unique attribute class taking class labels as its value. The set of attributes is expressed as At = F U {class}, where F is the set of attributes used to describe the objects and {class} is the decision attribute. The goal is to find rules from existing training dataset in the form of$\phi \Longrightarrow$ class =c_i, where $\phi$is a formula over F and $c_i$is a class label. In this paper, four classes based on the occurrences of earthquake during 20th century and the last six months are considered as values of the decision attribute. For simplification, four considered classes are labelled as below:

A= Earthquake with Magnitude more than 6 happened during 20th century
B= Earthquake with Magnitude less than 6 happened during 20th century
C= No occurrence of earthquake
D= Earthquake with Magnitude between 3 and 4 during the last six months

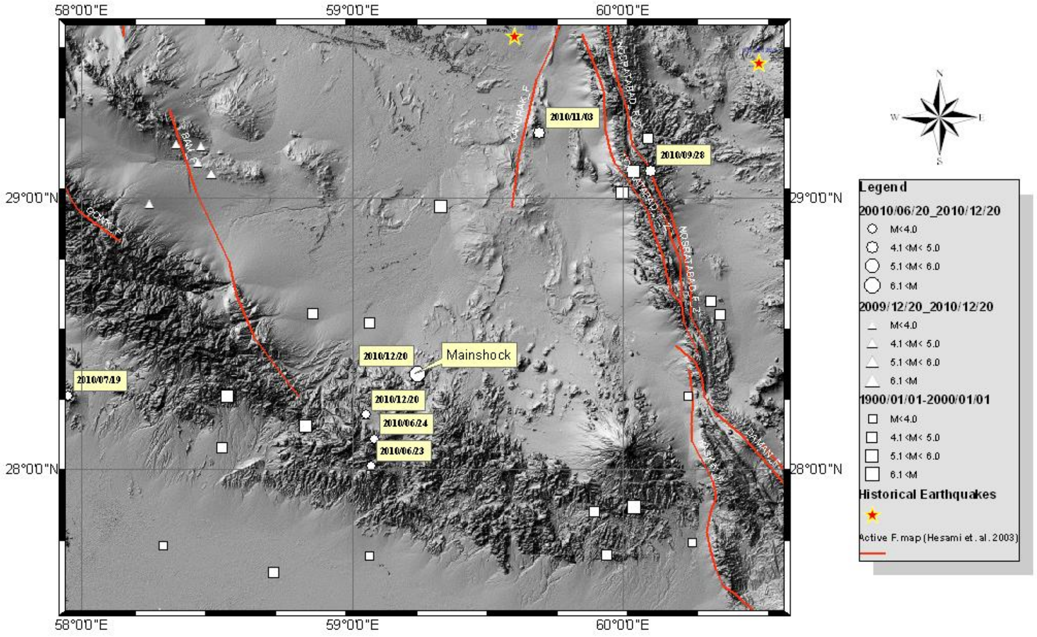

Figure1.Seismotectonic map of the studied area. Earthquakes are categorized based on different time periods. The date of events which have happened in the period of 2009/12/20-2010/12/20 are presented on the map. The historical events are shown by star symbol. The active fault map is from Hessami et. al., 2003.

Table1 illustrates the information table used in this research. For simplification, the titles of all criteria are given as follows:

Seis_Rate	The rate of seismic risk of the area
Fault	The existence of fault in the area and its length (L)
Fault_coin.	Fault coincidence in the area
More6.	Occurrence of historical earthquake with magnitude more than 6 Richter
Bet4_6.	Occurrence of historical earthquake with magnitude between 4 and 6 Richter in previous decade

CONCEPT FORMATION

The formation and representation of granules deal with algorithmic issues of granule construction. They address the problem of how to put two objects into the same granule. Aconcept which is definable in an information table is a pair of (ϕ, m (ϕ)), where ϕ∈ L. More specifically, ϕ is a description of m (ϕ) in S, i.e. the intension of concept (ϕ, m (ϕ)), and m (ϕ) is the set of objects satisfyingϕ, i.e. the extension of concept (ϕ, m (ϕ)). A formula has a meaning if it has an associated subset of objects.

To illustrate the idea developed so far, consider an information table given by Table 1. The following expressions are some of the formulas by the language L:

Fault =2A,

Fault_coin = nothing ∧ more6 = Event existence

The meanings of the above formulas are given by:

m (Fault =2A) = {u6}

m (Fault_coin = nothing ∧ more6 = Event existence) = {u14}

In these cases, the involved granules are {u6}, {u14}. To achieve these granules has not unique formulas. That is, there may exist two formulas such that m (ϕ) = m (ψ). For example; Fault =2A and Fault_coin = nothing∧ Fault =2A have the same meaning set {u6}.

One need to extract one of the possible solutions based on his/her understanding and preference. In the proposed algorithm, given a set of data, each user will try to make sense of data by observing it from different angles, in different aspects, and under different views. The preference can be stated order in granular computing. In the first step of the proposed algorithm, all possible granules from given database is constructed by an atomic formula which is given by $a = v$,where $a \in At$ and $v \in V_a$. Once the granules constructed, it is necessary to describe, to name and to label granules using certain languages. Also, a measurement is employed on each single granule that is named generality. It indicates the relative size of the granule. Indeed, a granule defined by the formula is more general if it covers more instances of the universe.

The quantity may be viewed as the probability of a randomly selected object satisfying the formula. Equation (2) shows the fraction of the size of a granule to the size of universe.

$$G(\phi) = \frac{|m(\phi)|}{|U|} \qquad (2)$$

Table 1. Information tableof the study area

Object	Seis_Rate	Fault	Fault_coin	More6.	Bet4_6.	Class
U1	Event existence	L>80	Nothing	Nothing	6Event	A
U2	Low occurrence- fault	L<2	Nothing	Nothing	1Event	C
U3	No event with fault in surrounding	L<2	Nothing	Nothing	1 Event	C
U4	Low occurrence- fault	3(L>80)& 20<L<80	2 Existence	Event existence	Nothing	A
U5	Low occurrence- fault	3(L>80)& 10<L<20	2 Existence	2 Event	Nothing	A
U6	Event existence	2(L>80)	Nothing	Nothing	7 Event	B
U7	Low occurrence- fault	L>80	Nothing	Nothing	2 Event	B
U8	No event with fault	20<L<80	Nothing	Nothing	Nothing	C
U9	No event with fault	3(L>80)	Nothing	Nothing	Nothing	B
U10	Low occurrence- without fault	L<2	Nothing	Nothing	Nothing	D
U11	Low occurrence- fault	L>80	Nothing	Nothing	Nothing	B
U12	Low occurrence- without fault	L<2	Nothing	Nothing	1 Event	D
U13	Noevent with fault in surrounding	4(L>80)	Existence	Nothing	Nothing	B
U14	Low occurrence- without fault	L<2	Nothing	Event existence	Nothing	B
U15	Low occurrence-fault	L<2	Nothing	Nothing	Nothing	B
U16	Event existence	L<2	Nothing	Nothing	Nothing	B
U17	No event with fault in surrounding	L<2	Nothing	Nothing	Nothing	B
U18	Low occurrence- fault	L>80	Nothing	Nothing	Nothing	B

RELATIONSHIP BETWEEN THE CONCEPTS

Based on the notions introduced so far, data mining for rules can be viewed as searching for relationship between the overlap concepts. By expressing rules with intensions of concepts, we may easily explain them in a natural language, provided that we can explain formulas of the language L. Therefore, a crucial issue is the characterization, classification, and interpretation of rules. It is reasonable to expect that different types of rules represent different kinds of knowledge derivable from a database. To achieve this goal, different quantitative measures and different mining algorithms can be designed. In many studies of machine learning and data mining, a rule is usually paraphrased by an “IF - THEN” statement, “if an object satisfies ϕ then the object satisfies ψ”. The interpretation suggests a kind of cause and effect relationship between ϕ and ψ. Basically, another issue to construct rules from input dataset is to identify relationship between two granules and relationship between a granule and a family of granules. So, to induct more confident relationship between existence concepts, three quantitative measures are implemented in this research as follow:

i) Confidence or absolute support: It is a measure of the correctness or the precision of the inference. The quantity can be computed by a fraction of number of samples in a granule that is equal to objects belong in a class, to the size of the granule. If the quantity of the confidence of a rule is kept high, then less number of association rules will be mined but their prediction accuracy will be quite high.

ii) Coverage: It is a measure of the applicability or recall of the inference and indicates fraction of data in a class correctly classified by the atomic formula. The quantity can be computed by fraction of number of samples that satisfy the THEN part of the rule, to the size of data with the same class label as the rule consequent.

iii) Conditional entropy: it provides a measure that is inversely related to the strength of the inference. This measurement depends on the confidence and one can identify one equivalent class in which the object belongs with no measure of uncertainty where an object satisfies the formula of attribute-value. In this case, confidence of the formula for at least one equivalent class is1.

INDUCTION OF MORE CERTAIN RULES

A granule network has |At| levels at most which consist of granules as its node and attribute-values as its branches. Granules in different levels are linked together by a relation in a hierarchy form. In the proposed top-down construction algorithm, a granule in a higher level can be decomposed into many granules in a lower level. A granule in a lower level provides detailed description of the granule in a higher level, and a granule in a higher level has a more abstract description than the granules in a lower level.

To construct a granule network it is required that first dividing the universe into grouping or partitions of the same class with atomic formula of attribute-values. In the proposed algorithm, three mentioned measures on relationship between the concepts are applied automatically. In this waymore certain pair attribute-value is selected at each step. The algorithm will continue until all objects are correctly classifiedin which, each object is associated with a unique class label.It is importantto find a subset of attribute-value with high coverage, confidence and minimumentropy. In this paper, construction of the tree are continued until all granules at the end level reach to zero entropy in which, union of all non-active granules be equal to the universe set.

The active node and non-active node at one level should be characterized to determine whether dividing of a granule should continue or not. A granule is considered non-active if it includes the two following conditions: (i) the granule to be a subset of unique class and (ii) union of all granules at low level and non-redundant covers the solution of the root granule. An active granule is further divided through efficient measures. After union of all non-active granules constructed from a covering solution of the universe set, construction of the decision granule tree would be stopped. This can identify the information that is providedby the constructed tree in the form of “IF - THEN” statements.

The proposed algorithm which includes the eight steps is presented as follows:

1: **Load** a dataset for classification.

2: **Set** U as the root node of granule tree at the initial stage.

3: **Construct** the family of basic concepts with respect to atomic formulas:

$$BC\ (U) = \{(a = v,\ m\ (a = v)) \mid a \in C,\ \ v \in V_a\}$$

4: **Compute** four parameters (Generality, Coverage, Confidence, Entropy) for each granule

5: **Se**t the granule network to GN = ({U},∅), which is a graph consisting of only one node and no arc.

6: **Set** the activity status of U.

7: **Select** the BC = (a = v, m (a = v)) with maximum value of fitness with respect to U.

8: **While** the set of non-active nodes is not a non-redundant covering solution of the consistent classification problem, do:

- 8-1: Select the active node N with the maximum value of activity (maximum entropy, minimum coverage).
- 8-2: Compute entropy and generality for the active nodes with respect to all remained attribute-values
- 8-3: Select the basic concepts with minimum intersection (the least overlap) with the union of all non-active nodes in the granule tree.
- 8-4: Select the basic concepts BC = (a = v, m (a = v)) with the minimum entropy among granules in 8-3.
- 8-5: If there is more than one concepts selected in 8-4, the one with the maximum sum of coverage of these concepts with respect to all the considered classes is selected.

8-6: Modify the granule network GN by adding the granule N ∩ m (a = v) as a new node, connecting it to N by arc, and labeling it by a = v.
8-7: Set the activity status of the new node.
8-8: Update the activity status of N.
9: **Export** a granule tree and its corresponding rules.

In this research, from 18 seismic areas, 36 different nodes were added to the granule network tree in the 5 levels. Among these extracted nodes, based on non-redundant covering solutions and user preferences, a user can freely explore the dataset according to his/her preferenceand priority. It is important to note that user can process various skills, intelligence, cognitive styles, frustration tolerances and other mental abilities. To give an example from the extracted rules in this part of research, 4 different rules which classified same objects are presented below;

If Fault=3(L>80) → {u9} (class B)
If Fault_coin= Nothing and Fault= 3(L>80)) → {u9} (class B)
If Fault_coin= Nothing and More6 = Nothing and Fault=3(L>80) → {u9} (class B)
If Fault_coin= Nothing and More6 = Nothing and Fault= L<2 and Bet4_6= Nothing and Fault=3(L>80) → {u9} (class B)

However, all of these kinds of rules have minimum uncertainty and provide consistent classification the first rule is considered more general than others. That is, it gives general information. But in some solutions, one needs to examine the problem at a finer granulation level with more detailed information when there is a need or benefit fordoing so.

CONCLUSIONS

This paper has proposed a new approach to induct the classification rules with less uncertainty for seismic risk assessment. The extracted rules with zero entropy and high accuracy can be evaluated by experts and selected based on their needs and the existence situations. These inducted rules and their information along may be used for prediction of the next events with near or same circumstances based on the rule accuracy. This paper is the result of part of the researches which are examined to find methods for knowledge representation, searching, and reasoning. It is the first step to use broader meaning for hierarchies, instead of the restricted mathematical notion defined by a partial ordering in which this paper shows the combination of the theory of hierarchy and the systems thinking, as well as taking advantages of both.

REFERENCES

Bargiela A and Pedrycz W (2002) Granular Computing: an Introduction, Kluwer Academic Publishers, Boston

Demri S and Orlowska E (1998) Logical analysis of indiscernibility, in: IncompleteInformation: Rough Set Analysis, Orlowska, E. (Ed.), Physica-Verlag, Heidelberg, pp.347-380

Hessami K, Jamali F and Tabasi H (2003) Major active fault of Iran. 1:2,500,000, Tehran, International Institute of Earthquake Engineering and Seismology

Hobbs JR (1985) Granularity, Proceedings of the Ninth International Joint Conference on Artificial Intelligence, pp.432-435

Inuiguchi M, Hirano S and Tsumoto S (2003) Rough Set Theory and Granular Computing, *Springer*, Berlin

Lin TY (1997) Granular computing, Announcement of the BISC Special Interest Group on Granular Computing

Lin TY, Yao YY and Zadeh LA (2002) Data Mining, Rough Sets and Granular Computing, Physica-Verlag, Heidelberg

Marr D (1982) Vision, A Computational Investigation into Human Representation and Processing of Visual Information, Freeman WH and Company, San Francisco

Moore RE (1966) Interval Analysis", Prentice-Hall, Englewood Cliffs, N.J

Pawlak Z (1982) Rough sets, *International Journal of Computer and Information Sciences*, 11, pp. 341-356

Pawlak Z (1991) Rough Sets: Theoretical Aspects of Reasoning about Data", Kluwer Academic Publishers, Dordrecht

Pedrycz W (2001) Granular Computing: An Emerging Paradigm, Physica-Verlag, Heidelberg

Samadi Alinia H, Delavar MR and Yao YY (2010) Support and confidence parameters to induct decision rules to classify Tehran's seismic vulnerability,*Proc. 6th International Symposium on Geo-information for Disaster Management (Gi4DM),* Sep. 15, Torino, Italy

Samadi Alinia H (2010) Assessment of Vulnerability of Earthquake in Tehran Using Granular Computing Model, MS.c. Thesis. (In Persian with English abstract), College of Engineering University of Tehran

Samadi Alinia H and Delavar MR (2010) Granular computing model for solving uncertain classification problems of seismic vulnerability , in Spatial Data Quality from Process to Decision, Edited by RodolpheDevillers and Helen Goodchild., CRC Press, Chapter12, pp.132-134

Shafer G (1976) A Mathematical Theory of EvidencePrinceton University Press: Princeton NJ, Wang G, Liu Q, YaoYY and Skowron, A (2003) Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing", LNCS 2639, Springer, Berlin.

Wille R (1992) Concept lattices and conceptual knowledge systems, Computers Mathematics with Applications, 23, pp.493-515

Yao YY and Zhong N (1999) Potential applications of granular computing in knowledge discovery and data mining, Proceedings of World Multiconference on Systemics, Cybernetics and Informatics, pp.573-580

Yao YY (2000) Granular computing: basic issues and possible solutions, *Proceedings of the 5th Joint Conference on Information Scie*nces, pp.186-189

Yao YY (2001) On Modelling data mining with granular computing, Proceedings of COMPSAC, pp.638-643

Yao JT and Yao YY (2002) Induction of Classification Rules by Granular Computing , *Proceedings of the Third International Conference on Rough Sets and Current Trends in Computing*, Lecture Notes in Artificial Intelligence, 331-338

Yao YY (2004a) Granular computing, *Computer Science (JiSuanJiKeXue)*, 31, 1-5

Yao YY (2004b) A partition model of granular computing, LNCS Transactions on Rough sets, Vol. I, pp. 232-253

Yao YY (2005) Perspectives of granular computing, *Proceedings of 2005 IEEE International Conference on granular computing*, Vol. 1, pp.85-90

Yao JT, Yao YY and Zhao Y (2005) Foundations of classification, in: Lin TY, Ohsuga S, Liau CJ and Hu X (Eds), Foundations and Novel Approaches in Data Mining, Springer, Berlin, pp. 75-97

Zadeh LA (1997) Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets and Systems, 19, 111-127

Zadeh LA (1998) granular computing and their roles in the conception , Some reflections on soft computing design and utilization of information/intelligent systems, Soft Computing, Vol. 2(1), pp.23-25

Zhang B and Zhang L (1992) Theory and Applications of Problem Solving, North-Holland, Amsterdam

Zhang L and Zhang B (2004) the quotient space theory of problem solving, Fundamental Informatcae, 59, pp.287-298

Zhao Y and Yao YY (2005) Interactive Classification Using a Granule Network. *Proceedings of the Fifth International Conference of Cognitive Informatics* (ICCI'05), pp.250-259

Zhao Y, Yao YY and Yan M (2007) ICS: An Interactive Classification System, Proceedings of the 20th Canadian Conference on Artificial Intelligence (CAI'07), pp.134-145

USGS (2004) http://earthquake.usgs.gov/recenteqsww/Quakes/uscvad.htm